Current Trends in Image Similarity Search and application in Fashion based Image Search Engine

HARSHIT AGARWAL*, Kalinga Institute of Industrial Technology, India

In this paper we have studied various traditional and current trends in use of artificial intelligence in Image Similarity Search. It explores and implements multiple state-of-the-art methodologies to develop an alternative to Google Lens, focusing on fashion-specific applications like cosplay recognition. The proposed approaches encompass feature extraction using pre-trained convolutional neural networks (ResNet, VGG, EfficientNet), deep metric learning through Siamese Networks and Triplet Loss, embeddings generated by Vision Transformers (ViTs) and CLIP, hashing-based retrieval methods such as Locality Sensitive Hashing (LSH), and latent space mappings via autoencoder architectures.

To train and evaluate these methods, CalTech101 dataset is used. Along with that, a custom cosplay image dataset was created by scraping a wide range of cosplay pictures from the internet. Images were categorized into separate folders representing distinct classes, enabling robust training of the different models. Fine-tuning and performance evaluation of the models were conducted using metrics such as precision, recall, and retrieval accuracy, with particular emphasis on computational efficiency and scalability for real-time usage scenarios.

Through comparative analysis, this study identifies the strengths and limitations of each technique, providing insights into their suitability for various use cases in fashion-focused image search, from high-speed retrieval systems to scenarios demanding high accuracy and robustness. The findings highlight effective solutions for advancing image similarity search technologies in fashion and related domains.

Additional Key Words and Phrases: Image Similarity Search, Deep Metric Learning, Vision Transformers, CBIR, Resnet, fashion, cosplay

ACM Reference Format:
Harshit Agarwal. 2024. Current Trends in Image Similarity Search and application in Fashion based Image Search Engine. 1, 1 (December 2024), 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

1.1 Background

Introducing Image Similarity Search, the problems of Image Retrival on the basis of Image Content, Product etc.

1.2 Objective

The aim of this paper is to study the various methods of image similarity search or CBIR (Content based Image retrieval) for specific use cases and their performances.

Author's Contact Information: Harshit Agarwal, 23052801@kiit.ac.in, Kalinga Institute of Industrial Technology, Bhubaneshwar, Odisha, India.

ACM XXXX-XXXX/2024/12-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 Literature Review

As noted in the introduction, this report covers a wide range of models and techniques for the problem of image similarity search.

In a general CBIR pipeline, the images are segmented and features are extracted from them using methods like SURF,FAST,ORB etc. and then a Nearest Neighbor search is trained on them.

Feature extraction using pre-trained convolutional neural networks (CNNs) like ResNet[4], VGG[14], or EfficientNet[15], followed by nearest neighbor search (e.g., k-NN or cosine similarity).

Deep metric learning approaches such as Siamese Networks or Triplet Loss-based models[6]

Visual embeddings generated via Vision Transformers (ViTs)[3] or CLIP[12].

Hashing-based methods such as Locality Sensitive Hashing (LSH) or deep learning-based hashing [5]

Autoencoder-based image reconstruction to map images into a latent space for similarity comparison. [17]

Other papers/models read: [1] [2] [3] [4] [5] [6] [7] [10] [8] [11] [12] [13] [14] [15] [16] [17] [18]

3 Methodology

3.1 Dataset

In this paper we primarily run tests on the Caltech-101 dataset.

The Caltech-101 dataset [9] has a background class, but the background class of that dataset contains a few distinctive images. Moreover, several background images of that dataset contain human faces that overlap with another class. Therefore, we removed the background class when predicting and training using our model.

However, we also curate a dataset of our own by scraping images of cosplays from a custom google search engine and storing them as classes of their own.

3.2 Architecture

In the following report we have implemented and tested the following models falling under various different categories.

3.2.1 Traditional Features. Features like SIFT, SURF and ORB are extracted for all the images as a feature list. After that, K-Nearest Neighbour with KDTree and L2 are trained on the feature list.

3.2.2 Pre-trained Deep learning Models. Models like ResNet, VGG, MobileNet and EfficientNet are used to extract and store a feature list of the dataset images. After that, K-Nearest Neighbour with KDTree and L2 are trained on the feature list.

3.2.3 Pre-trained Vision Transformers. Models like Google's Vision Transformer, CLiP are used to extract vision features of the training

images and stored in a feature list. After that, K-Nearest Neighbour with KDTree with euclidian distance metrics are trained on the feature list.

3.2.4 Deep Metrics. Models with the methodology of a Siamese Network with Triplet Loss is used to extract vision features of the training images and stored in a feature list. After that, K-Nearest Neighbour with KDTree and L2 are trained on the feature list.

All of these models were coded on Google Colab and feature extractions as well as training was done on Colab's T4.

For the cosplay datasets, a Dense layer is added on the Pretrained models. And the models are retrained.

For the Resnet, this is the parameters in the model and this (1) is the training curve of it.

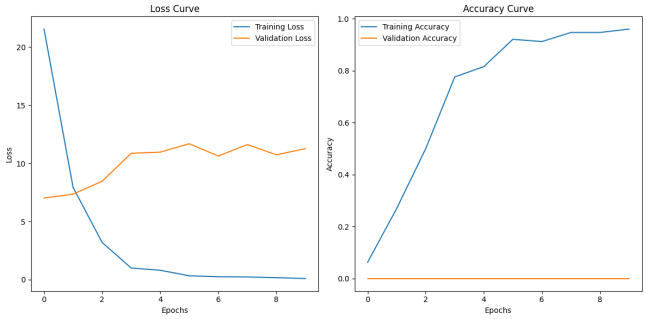

Fig. 1. Training curve for Resnet retraining

3.3 Testing

To test the models, the Caltech 101 dataset is segmented using a text classifier model to classify the different classes into parent classes.

Parent class includes "Animals, Miscellaneous, Electronics,

Out of every child class, we have extracted 3-4 images at random and made predictions on them. See tables, 1 2

Similarly, for every parent class, we have generated all the scores for all the models as well. See tables, 3

During the predictions we keep track of metrics like nDCG, Recall, mAP (Precision) and Retrieval Time of the models predictions and the average scores for the models are reported for the overall dataset.

The second curated dataset of cosplay images is then tested using a similar fashion. Images are loaded randomly and average metrics are reported in 6

4 Result

Table 1. Performance Metrics for Image Similarity Models based on K-NN

Model	Precision	Recall	nDCG	Retrieval Time (s)
ResNet	0.84	1.00	0.93	0.10
EfficientNet	0.90	1.00	0.95	0.07
VGG	-	-	-	-
MobileNet	0.41	1.00	0.56	0.18
ViT	0.95	1.00	0.98	0.88
DINOv2	0.96	1.00	0.98	0.03
CLiP	0.89	1.00	0.95	0.02
ViT-1.58b	-	-	-	-
Siamese Network	-	-	-	-
SN	-	-	-	-
Autoencoder	-	-	-	-

Model training time: 30 Mins for both models.

Table 2. Performance Metrics for Image Similarity Models based on Local Sensitive Hashing

Model	Precision	Recall	nDCG	Retrieval Time (s)
ResNet	0.84	1.00	0.93	0.10
EfficientNet	0.90	1.00	0.95	0.07
VGG	-	-	-	-
MobileNet	0.41	1.00	0.56	0.18
ViT	0.95	1.00	0.98	0.88
DINOv2	0.96	1.00	0.98	0.03
CLiP	0.89	1.00	0.95	0.02
ViT-1.58b	-	-	-	-
Siamese Network	-	-	-	-
SN	-	-	-	-
Autoencoder	-	-	-	-

Table 3. Performance Metrics for Classes based on Resnet Model

Class	Precision	Recall	nDCG	Retrieval Time (s)
Miscellaneous	0.86	1.0	0.94	0.30
Animals	0.81	1.0	0.91	0.36
Electronics	0.93	1.0	0.97	0.28
Vehicles	0.91	1.0	0.96	0.28
Nature	0.90	1.0	0.95	0.27
Furniture	0.80	1.0	0.918	0.24
Clothes	0.96	1.0	0.99	0.37
Appliances	0.96	1.0	0.99	0.23

Table 4. Performance Metrics for Classes based on Mobilenet Model

Class	Precision	Recall	nDCG	Retrieval Time (s)
Miscellaneous	0.65	1.0	0.78	0.17
Animals	0.22	1.0	0.35	0.12
Electronics	0.49	1.0	0.64	0.12
Vehicles	0.53	1.0	0.65	0.11
Nature	0.38	1.0	0.52	0.15
Furniture	0.71	1.0	0.86	0.13
Clothes	0.63	1.0	0.73	0.15
Appliances	0.56	1.0	0.79	0.11

Table 5. Performance Metrics for Classes based on ViT Model

Class	Precision	Recall	nDCG	Retrieval Time (s)
Miscellaneous	0.95	1.0	0.98	0.92
Animals	0.90	1.0	0.95	1.21
Electronics	0.97	1.0	0.99	0.95
Vehicles	0.96	1.0	0.98	0.99
Nature	0.92	1.0	0.96	0.92
Furniture	0.96	1.0	0.98	0.97
Clothes	0.87	1.0	0.95	0.93
Appliances	0.93	1.0	0.97	1.11

Table 6. Performance Metrics for Image Similarity Models on Cosplay Dataset based on K-NN

Model	Precision	Recall	nDCG	Retrieval Time (s)
ResNet	0.41	1.00	0.73	0.28
EfficientNet	0.90	1.00	0.95	0.07
VGG	-	-	-	-
MobileNet	0.41	1.00	0.56	0.18
ViT	0.95	1.00	0.98	0.88
DINOv2	0.96	1.00	0.98	0.03
CLiP	0.89	1.00	0.95	0.02
ViT-1.58b	-	-	-	-
Siamese Network	-	-	-	-
SN	-	-	-	-
Autoencoder	-	-	-	-

Model training time: 30 Mins for both models.

5 Discussion

dcdscds

6 Future Directions

dsdss

7 Conclusion

sfdsfs

References

[1] Ahmad Anis. 2023. Diving into clip by creating semantic image search engines. https://medium.com/red-buffer/diving-into-clip-by-creating-semantic-image-search-engines-834c8149de56

[2] Tapan Babbar. 2024. Build an ai image similarity search with transformers — vit, clip, dino-v2, and blip-2. https://medium.com/@tapanbabbar/build-an-image-similarity-search-with-transformers-vit-clip-efficientnet-dino-v2-and-blip-2-5040d1848c00

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: transformers for image recognition at scale. https://doi.org/10.48550/arXiv.2010.11929 arXiv:2010.11929.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, 770–778. https://doi.org/10.1109/CVPR.2016.90

[5] Sarthak Joshi. 2023. Understanding locality sensitive hashing(Lsh): a powerful technique for similarity search. https://medium.com/@sarthakjoshi_9398/understanding-locality-sensitive-hashing-lsh-a-powerful-technique-for-similarity-search-a95b090bdc4a

[6] Kassem. 2024. Image similarity estimation using a siamese network with triplet loss: a practical guide. https://elcaiseri.medium.com/image-similarity-estimation-using-a-siamese-network-with-triplet-loss-a-practical-guide-124938e24b3a

[7] Simon Lepage, Jérémie Mary, and David Picard. 2024. Lrvs-fashion: extending visual search with referring instructions. https://doi.org/10.48550/arXiv.2306.02928 arXiv:2306.02928.

[8] Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. 2022. Caltech 101. https://doi.org/10.22002/D1.20086

[9] Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. 2022. Caltech 101. https://doi.org/10.22002/D1.20086

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. https://doi.org/10.48550/arXiv.2301.12597 arXiv:2301.12597.

[11] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. Dinov2: learning robust visual features without supervision. https://doi.org/10.48550/arXiv.2304.07193 arXiv:2304.07193.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. https://doi.org/10.48550/arXiv.2103.00020 arXiv:2103.00020.

[13] Yuki Shizuya. 2024. Vision embedding comparison for image similarity search: efficientnet vs. https://pub.towardsai.net/vision-embedding-comparison-for-image-similarity-search-efficientnet-vs-4eac6bf553c4

[14] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. https://doi.org/10.48550/arXiv.1409.1556 arXiv:1409.1556.

[15] Mingxing Tan and Quoc V. Le. 2020. Efficientnet: rethinking model scaling for convolutional neural networks. https://doi.org/10.48550/arXiv.1905.11946 arXiv:1905.11946.

[16] UATeam. 2024. Understanding image similarity with machine learning. https://medium.com/@aleksej.gudkov/understanding-image-similarity-with-machine-learning-c8680c24dd56

[17] Dagang Wei. 2024. Demystifying neural networks: similar image search with autoencoder. https://medium.com/@weidagang/demystifying-neural-networks-similar-image-search-with-autoencoder-d15eedbae436

[18] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (Dec. 2014), 67–78. https://doi.org/10.1162/tacl_a_00166